**Impact of Educational Attainment on Wages**

**Anri Sorel and Erin Shinners**

**ECON 3161: Econometric Analysis**

**Dr. Shatakshee Dhongse**

**Fall 2019**

**Abstract:**

Research in labor economics often tests the effect of educational attainment on wages. Our paper examines this relationship using individual-level data from the 2017 American Community Survey Public Use Microdata Sample for the state of Georgia. Across three regression models, we select four additional test variables: age, English fluency, race, and gender. We hypothesize that educational attainment will be the key determinant in wages. The results of our regression models did not fully support our hypothesis, with the explanatory variable gender having the most significant effect.

## I. Introduction

The effect of educational attainment on wages and subsequent economic growth is a widely tested model in U.S. labor economics. The importance of studying education cannot be understated, as it has a strong correlation to different facets of development and financial well-being on the individual, state, and global levels. Rising costs for primary, secondary, and tertiary education, coupled with an increasingly competitive and volatile labor market, provide an environment continually and inherently curious about the relationship between educational attainment and subsequent wages upon employment. This relationship is of utmost financial importance to individuals debating the cost-benefit analysis of investing in current or future schooling. Additionally, administrators in education and policymakers must keep this relationship in mind when marketing and budgeting for educational programs. Using cross-sectional data from the 2017 Public Use Microdata Survey conducted by the American Community Survey on individuals in the State of Georgia, our study examines the impact of educational attainment on wages adjusted for inflation.

It is classically accepted that a productive worker sees the highest returns in the form of wages. The human capital theory adopted in the mid-twentieth century gave a name to the categories and variables responsible for increasing worker productivity. One of the categories, economic capital, consists of traits possessed by humans that allow them to contribute to their personal economic value. Education is a key trait in economic capital believed to increase a worker's overall productivity and marketability to employers. Naturally, this is associated with an increase in wages. With the general acceptance of the human capital theory coupled with the questions of the stakeholders mentioned above (potential students, administrators, and policymakers), there has been a notable increase in the number of econometrics studies testing the correlation of educational attainment on wages. In decades past in the United States, exhibiting higher levels of education was often associated with the promise of more gainful employment. Previous economic models traditionally support this view. Today, there is an increase in educational attainment. However, the prospect of yielding a high wage, or for gaining employment at all, is societally thought to have diminished significantly with the preponderance of other determining variables.

In this study, we hypothesize that despite the dismal current social convention, educational attainment continues to be the key determinant of wages as predicted by the human capital theory. In our simple regression model, we expect the independent variable educational attainment to be positively correlated to the dependent variable wage. For our multiple regression model, we expanded upon education and selected additional independent variables for testing: age and ability to speak English.

While we expect the multiple regression model to also yield a positive correlation between the newly introduced independent variables, we maintain our hypothesis that educational attainment yields the greatest positive impact on worker's wages.

## II. Literature Review

Literature about labor economics and education frequently touches on the Mincer equation. Traditionally, the relationship between educational attainment and wages is referred to as the "Mincer rate of return," after the economist who popularized it, Jacob Mincer. Patrinos (2016) proposes this model as an estimate on the individual private monetary return in the form of wages for additional years of attained education. After tabulating the Mincerian rate of return to education for 136 economies, Patrinos's (2016) reports a result in alignment with our hypothesis that returns to education are positive. Patrinos (2016) contributes to wage-education literature, stating that when considering the return on education, one must also consider the statistical significance of years of experience as an explanatory variable due to its strong correlation to an increased human capital that yields higher wages. Failure to include experience in the model triggers the omitted variable bias and results in an underestimation of the impact of education on wage. After totaling the surveyed economies, the average return to education is 9.7%. His work takes the model a step further, adding in categorization by gender, showing that female returns to education are always higher.

Psacharopoulos and Patrinos (2004) discuss a general overview of the return to education until 2004 and offer a critique of traditional sampling methodology. They conclude, interestingly, that while returns on education are positive, they are falling. From 1992 to 2004, the return on education fell 0.6 percentage points, despite an increase in the supply of educational programs and an increase in enrollment. This confirms the social convention that is held today that motivated our topic: one may be more educated in the modern market, but their wages aren't necessarily comparable to those of your equal in decades past. Psacharopoulos and Patrinos (2004) also touch on the critiques of traditional regression studies of education and wage. Traditional studies draw on data compiled using a survey of firms, which results in bias thanks to the nature of the study. Surveys of firms are often skewed towards the urban working class, which is not representative of any single country's true population. Psacharopoulos and Patrinos (2004) state that surveys of households are preferred when regression education on wages. This is encouraging for our results, given our data source, the American Community Survey Public Use Microdata Sample, draws on both household and individual data across the State of Georgia.

Psacharopoulos and Patrinos (2018) provide a more concise overview of twenty-first-century data on returns to education in their decennial review. This study expands on their 2004 update, primarily

reinforcing the idea that across 705 estimates from 1950 to 2014, the private returns to education remain positive yet declining. This study adds that while falling, the decline is slight, with the 2004 average rate yielding a return of 9.7% and the 2018 average rate yielding a return of 9.5%. Psacharopoulos and Patrinos (2018) conclude that this decline is insignificant, and the result of an environment where the cost of education remains the same or increases despite an increase in supply. They illustrate this idea further by focusing on a multitude of individual countries across a timespan of decades.

Card (2001) explores the wage-education model using a slightly different perspective than that of other literature. He touches on literature from the late twentieth century that states returns on education are positive with respect to wage; however, the model should be evaluated on a case by case basis where education is optimized with special attention paid to the costs of attending school and not simply by paying attention to highest wage. Card (2001) also elaborates that "individuals may have different aptitudes and tastes for schooling relative to work," and as such, educational attainment should be selected with one's future industry in mind. Throughout his work, Card (2001) reiterates the importance of considering an individual's environment when evaluating an education-wage equation, particularly one's incentivization to yield a higher wage and the marginal cost to attain additional education. In the instance of excluding both incentivization and the marginal cost of education, Card (2001) states that the OLS regression estimate of education on wages is biased. While the literature is dated, this conclusion is important to consider in our attempt to uncover the relationship between wages and education in our data while still considering the possibility of unknown variables in the error term.

Our study hopes to contribute to the existing literature with similar results. Per our hypothesis, we expect our explanatory variable educational attainment to yield a higher wage and to hold as the strongest variable in a multiple regression model against other variables of interest. Additionally, we hope to expand on the literature by diversifying our additional variables in the multiple regression model to include the age of the individual and the worker's English fluency. When comparing our results to that of previous literature, it is important to remember the source of our data. While we expect similar results, our models are based on data specific to the state of Georgia. Also, our data will examine a single year in time to give more concise and modern results, as opposed to a year-by-year comparison.

## III. Data

### A. Source of Data

We drew our variables from the American Community Survey Public Use Microdata Sample. The Public Use Microdata Sample draws a sample from the aggregate Census data to allow for easy customization and tabulation. For our data, we selected the individual Public Use Microdata Sample,

meaning the variables report on single people and not households. Additionally, we selected the year 2017, a 1-year sample, which is representative of 1% of the United States population. For reporting purposes, we decided to focus on the state of Georgia for relevance to the expected audience. The regression was calculated using anywhere from approximately 4,000 to 6,000 observations.

### B. Description of the Variables

| Variable | Source | Year | Observation* | Type |
|---|---|---|---|---|
| *logwagp* | ACS PUMS | 2017 | 47,795 | Dependent |
| *schl* | ACS PUMS | 2017 | 96,647 | Independent |
| *agep* | ACS PUMS | 2017 | 81,031 | Independent |
| *eng* | ACS PUMS | 2017 | 11,082 | Independent |

Figure 1 *The difference in observations is discussed in Part C (Descriptive Statistics of the Variables)

The explained variable to be tested is the natural log of the population wage (*logwagp*). This variable specifically records an individual's wages or salary income in the past 12 months adjusted into constant dollars to account for inflation. A natural log was selected to account for the wide range of values for population wage compared to the range of other recorded variables (scalability). Additionally, the natural logarithm of *wagp* allows for the results of the independent variables to be interpreted as the percentage change in *wagp* when a given independent variable increases by one. The primary explanatory variable to be tested is the level of educational attainment (*schl*). Further explanatory variables include age (*agep*) and English proficiency (*eng*).

The inclusion of educational attainment is to test the hypothesis that the wage a given worker receives is influenced by his total level of education. The other independent variables, age (*agep*) and English proficiency (*eng*) are included in order to maintain the assumption of *ceteris paribus.* It is reasonable to assume that demographic factors other than education determine overall wage level. In order to minimize the value of the error term and ensure the data are not underdefined, one must measure the change in wage from a change in education when all other possible influences are held constant. The secondary independent variables (*agep*, *eng*) are included to mitigate the possibility of underdefined data. Including age (*agep*) accounts for possible wage differences relating to age, like possible seniority. Including English proficiency (*eng*) accounts for wage differences, possibly due to language barriers preventing higher wage attainment in a state where business is predominately conducted in English. Based on the following coding for *eng*, an increase in the variable equates to a decrease in proficiency, therefore, in this case, we expect the relationship between *logwagp* and *eng* to be negative.

When interpreting the variables for educational attainment and English proficiency, consider the following unique coding scales given by the American Community Survey Data Dictionary:

*schl:* 01 = No schooling completed; 02 = Nursery school, preschool; 03 = Kindergarten; 04 = Grade 1; 05 = Grade 2; 06 = Grade 3; 07 = Grade 4; 08 = Grade 5; 09 = Grade 6; 10 = Grade 7; 11 = Grade 8; 12 = Grade 9; 13 = Grade 10; 14 = Grade 11; 15 = 12th grade - no diploma; 16 = Regular high school diploma; 17 = GED or alternative credential; 18 = Some college, but less than 1 year; 19 = 1 or more years of college credit, no degree; 20 = Associate's degree; 21 = Bachelor's degree; 22 = Master's degree; 23 = Professional degree beyond a bachelor's degree; 24 = Doctorate degree

*eng*: 1 = Very well; 2 = Well; 3 = Not well; 4 = Not at all

### C. Descriptive Statistics of the Variables

The descriptive statistics of the variables in question are detailed in Figure 2.

| Variable | Observations | Mean | Std. Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| *logwagp* | 47,795 | 10.2 | 1.3 | 1.3 | 13.1 |
| *schl* | 96,647 | 15.9 | 5.6 | 1 | 24 |
| *agep* | 81,031 | 47.8 | 18.9 | 16 | 92 |
| *eng* | 11,082 | 1.6 | 0.8 | 1 | 4 |

*Figure 2*

While all data came from the same source, practical limitations prevent the number of observations from being identical for every variable. For *eng*, the original source of the data did not have the information available for every survey participant. As such, the original number of observations was lower. For other variables, the number of observations included in the analysis had to be pruned from the original data in order to better fit the research subject. For instance, the variable *logwagp* does not include any observations for which the observed wage is zero. This is not only a practical necessity of using a logarithm— the logarithm of zero is undefined — but eliminates any observations not pertinent to the research question. An individual with a wage of zero is not in the workforce and is not relevant to the research question. Additionally, individuals under the age of 16 were likewise omitted because they are not a part of the workforce.

To see the correlation between each of the independent variables and the dependent variable, see Appendix A, Figures 3, 4, and 5. These correlations are weak, yet seemingly positive for *logwagp* and *schl*, and *logwagp* and *agep*, and slightly negative for *logwagp* and *eng*. Due to the large range and number of observations, it is admittedly difficult to read the scatters.

### D. Checking the Gauss Markov Assumptions

1. Linear in Parameters — The model consists of a multiple linear regression that is linear in parameters and follows $Y = \beta_0 X_0 + \beta_1 X_1 + \ldots + u$.

2. Random Sampling — The original data source includes approximately 100,000 observations from within the state of Georgia, across a wide range of demographics. The sampling may be assumed to be random.
3. No perfect collinearity — There is no perfect collinearity between the variables in the model as determined by the following correlation matrix. The variables with the highest correlation were *agep* and *logwagp* as well as *schl* and *logwagp* , which was expected. The negative correlation for *eng* could be attributed to the way the variable is coded.

| | *logwagp* | *schl* | *agep* | *eng* |
|---|---|---|---|---|
| *logwagp* | 1.000 | | | |
| *schl* | 0.270 | 1.000 | | |
| *agep* | 0.329 | 0.057 | 1.000 | |
| *eng* | -0.120 | -0.483 | 0.108 | 1.000 |

*Figure 6*

4. Zero Conditional Mean — The expected value of the error term is zero. Realistically speaking, a value of zero cannot be achieved due to the limited information available. A large sample size and the inclusion of as many relevant independent variables as possible helps to minimize the value of the error term.
5. Homoscedasticity — This model is constructed with the assumption that a change in any one independent variable will not affect the value of the error term.

## IV. Results

### A. The Simple Regression Model

The equation for the simple regression model is as follows:

$$log(wagp) = \beta_0 + \beta_1 schl + u$$

$$log(wagp) = 7.845 + 0.126 schl$$

A complete view of the Stata output that obtained this model is under Figure 7 in Appendix A.

### B. The Multiple Regression Model

The equation for the multiple regression model is as follows:

$$log(wagp) = \beta_0 + \beta_1 schl + \beta_2 agep + \beta_3 eng + u$$

$$log(wagp) = 8.096 + 0.057 schl + 0.029 agep - 0.060 eng$$

A complete view of the Stata output that obtained this model is under Figure 8 in Appendix A.

### C. Interpretation of the Results

In order to determine the effects of the secondary explanatory variables on our explained variable, we performed a simple regression followed by a multiple regression analysis of the data. The simple linear regression analysis shows the gross effect an increase in educational attainment will have on the log of wages in the sample, without regard to the effects any other factors may have on wage. The simple regression has a larger sample size than the multiple regression due to the lower threshold for the required data. This provides a more random sample, which would distribute the data more closely to an approximately normal distribution. Other things equal, this would imply a model that more closely resembles the population.

However, while the simple regression strengthens the second Gauss-Markov assumption, it weakens the fourth. Without defining other parameters that can affect wage as independent variables, these parameters are included in the error term. This weakens the predictive ability of the model, as several factors that have a measurable effect on the explained variable are not being accounted for. Quantitatively, this weakening may be seen in the correlation coefficients of the two models.

In interpreting the coefficients of our model, we yield expected economic results. In the simple regression model, an increase in the education attainment variable, *schl*, yields a 12.6% increase in wages. This is in line with our hypothesis that educational attainment will positively impact wages. In the multiple regression model, we see that *schl* yields a 5.7% increase in wages, *agep* yields a 2.8% increase in wages, and *eng* yields a 6.0% decrease in wages. It may seem that *eng* does not follow our expectation that an increased proficiency equates to an increase in wages, however it is important to remember the coding of the variable provided by the ACS PUMS Data Dictionary. Since an increase in *eng* is coded to represent a lower proficiency, the multiple regression outcome is not completely out of line. In short, a decrease in English proficiency still equates to a decrease in wages.

Our multiple regression analysis showed a moderately stronger correlation compared to the simple regression, returning R-squared values of 0.173 and 0.105, respectively. While even the R-squared value of the multiple regression is comparatively low, explaining less than 20% of the variation in *logwagp,* the comparison of R-squared between the simple and multiple regression models yields a difference of approximately 7% demonstrates the utility of accounting for as many explanatory variables as possible. By adding age and English proficiency as independent variables, an additional percentage of the dependent variable's variance was explained. Going forward, we should consider minimizing the omitted variable bias by including more explanatory variables in future regression models.

For any given value of *schl* in the model, there are many observations, with a wide range of *logwagp*. Introducing new independent variables would significantly reduce the unexplained variation and bias and hopefully would increase the R-squared value. However, due to the preponderance of data-less factors that may influence wage, such as IQ or familial connection, accounting for every possible explanatory variable is not a realistic goal. As a result, there would be an upper limit on the R-squared value one could hope to achieve.

**D. Statistical Inference**

The multiple regression model yielding the following t-value, p-test, and confidence intervals:

| Independent Variable | t-value | P > \| t \| | 95% Confidence Interval |
|---|---|---|---|
| *schl* | 17.12 | 0.00 | 0.051— 0.064 |
| *agep* | 26.97 | 0.00 | 0.027— 0.031 |
| *eng* | -3.19 | 0.001 | -0.097— -0.023 |

*Figure 9; See Figure 8 for source of values*

Immediately, the high t-values and low p-values in *schl* and *agep* led us to believe that we could reject the null hypothesis (H0: Bj = 0) in favor of the alternative hypothesis. When tested against the absolute critical values at three different levels of significance, 1%, 5%, and 10%, and using 6,016 degrees of freedom, we found that both variables' absolute t-values were larger than the critical value, allowing us to indeed reject the null hypothesis. Both *schl* and *agep* are statistically significant in determining *logwagp.* The variable *eng* yielded a critical value of 2.576 at the 1% level of significance. Given that the t-value for *eng* was –3.19, one may conclude that it is statistically insignificant, however taking the absolute t-value means that we can reject the null hypothesis and conclude that English proficiency is indeed statistically significant to our model. The statistical inference conclusions of all three variables are supported by their low p-values. This is represented in the following table, Figure 10. Since each variable was individually significant at 1%, we will maintain this model going forward.

| Independent Variable | Model 1 | Model 2 |
|---|---|---|
| *schl* | 0.126***<br>(0.00169) | 0.057***<br>(0.003) |
| *agep* | | 0.029 ***<br>(0.001) |
| *eng* | | -0.060***<br>(0.019) |
| Intercept | 7.845<br>(0.03181) | 8.096<br>(0.084) |
| No. of obs. | 47795 | 6020 |
| R-square | 0.105 | 0.173 |

*Figure 10; Statistical Significance: 1%***, 5%**, 10%**

## V. Extensions

### A. Robustness Tests

#### #1: The F-Test

To test whether the variables *agep* and *eng* have a jointly significant effect on the explained variable *logwagp*, we performed an F-test with the unrestricted model depicted in Figure 8 and the restricted model depicted in Figure 7. The F-statistic was constructed from the ratio of:

- The difference of the residual sum of squares of the two models divided by the number of restrictions
- The unrestricted model's residual sum of squares divided by its degrees of freedom

With null hypotheses $\beta_1$, $\beta_2$=0. Using the *test* function in STATA, the F-statistic for the variables *agep, eng* was determined to be 364.16, with a p-value of 0.000. 364.16 being greater than 0, we reject the null hypothesis. Collectively, these variables do have a significant effect on the value of *logwagp*. This fact serves to reinforce the notion that accounting for these and other explanatory variables is necessary in order to construct an accurate model of wage determinants.

| *agep+eng* | |
|---|---|
| **SSR Unrestricted Model** | 7214.727 |
| **SSR Restricted Model** | 71034.902 |
| **Numerator Degrees of Freedom** | 2 |
| **Denominator Degrees of Freedom** | 6016 |
| **F-Statistic** | 364.16 |

*Figure 11*

#### #2: Multicollinearity

We tested for perfect collinearity of the independent variables in Section III, Part D (Checking the Gauss Markov Assumptions), Figure 6, and found none. When reexamining the table for multicollinearity between the independent variables, we find that *schl* and *eng* have the strongest relationship of –0.48 (a negative relationship). This could be attributed to the fact that those with poor English fluency are not able to obtain additional years of schooling in a predominately English education system. The remainder of the variables have weaker linear relationships; however, it is expected that the variables are somewhat related. For example, *agep* and *logwagp* have a weaker positive relationship, likely because an increase in age is generally correlated to a higher wage.

### B. Functional Form

The functional form of our model is a log-linear regression model. Our dependent variable, *logwagp*, was explained in terms of three different linear variables, *schl*, *agep*, *eng*.. The log-linear model allows us to account for any exponential relationship between the variables and provides scalability. Per the previous section, we saw that when holding all other factors equal, *schl* yields a 5.7% increase in wages, *agep* yields a 2.8% increase in wages, and *eng* yields a 6.0% decrease in wages.

**C. Dummy Variables**

Given the importance of each of the variables at a 1% level of significance, we decided to keep all the variables going forward. To reduce the omitted variable bias and enrich the model's explanation of wages, we added the variable *racwht*. This variable is also obtained from the 2017 ACS PUMS dataset and represents a dummy variable, where the base group is non-white respondents tested against white respondents. We chose to examine race because it is suspected of having many cultural biases that result in various life implications, such as wage, depending on the race observed. Since our data examines respondents in Georgia, we chose *racwht* as opposed to another race variable because the predominant race in the state is white, representing 60% of the population. We also chose to add a variable examining gender, since the gender wage gap is popularly discussed in media. For the sake of the model, we generated the variable *gender* from the existing variable *sex* and recoded the variable to represent a traditional binary dummy variable. The interpretation of the two dummy variables is as follows:

| Variable Name | Values | Source |
|---|---|---|
| *racwht* | 0 if non-white; 1 if white | ACS PUMS 2017 |
| *gender* | 0 if male; 1 if female | ACS PUMS 2017 |

*Figure 12*

The addition of the two dummy variables yielded the results shown in the equation below. The regression outputs are in Appendix A Figure 13.

$$log(wagp) = \beta_0 + \beta_1 schl + \beta_2 agep + \beta_3 eng + \beta_4 racwht + \beta_5 gender + u$$

$$log(wagp) = 8.232 + 0.062 schl + 0.028 agep - 0.064 eng + 0.026 racwht - 0.487 gender$$

The dummy variables showed a minimal change in the previous variable's economic significance. Since we are using *logwagp,* the dummy variable can be economically interpreted as the percent change in *logwagp* holding all other factors constant. Controlling for all other factors, *racwht* had a minimal effect on *logwagp* at 2.6%, similar to that of *agep*. The variable *racwht* had a significant economic effect

on *logwagp*, showing that females salaries are 48.7% lower when holding all else constant. Adding the two dummy variables also resulted in an increase in the R-squared value, to the tune of 4%. In short, adding the two variables achieved our goal of enriching the model's explanation of *logwagp,* however additional variables should be considered to increase R-squared in future models.

The dummy variables yielded the following t-value, p-test, and confidence intervals:

| **Independent Variable** | **t-value** | **P > \|t\|** | **95% Confidence Interval** |
|---|---|---|---|
| *schl* | 18.85 | 0.000 | 0.056 - 0.068 |
| *agep* | 27.42 | 0.000 | 0.026 - 0.030 |
| *eng* | -3.47 | 0.001 | -0.100 - -0.028 |
| *racwht* | 0.94 | 0.350 | -0.028 - 0.080 |
| *gender* | -17.42 | 0.000 | -0.540 - -0.431 |

*Figure 13; Source in Appendix A*

We can draw the same statistical conclusions about the variables *schl*, *agep*, and *eng* from Model 2. All of the variables are statistically significant at 1%, 5% and 10% using 6,014 degrees of freedom. We also reject the null hypothesis (H0: Bj = 0) in favor of the alternative for *gender* at all levels of significance, given that the absolute t-value is larger than the critical value at each level. For *racwht,* we fail to reject the null hypothesis at all levels of significance because the t-value, 0.94, is less than each critical value. The statistical significance for *schl, agep, eng,* and *gender* is supported by their low p-values, or the smallest level of significance at which the null hypothesis can be rejected. Consequently, the statistical insignificance of *racwht* is supported by its larger p-value. Additionally, the 95% confidence interval supports the same conclusions; since the value of our null hypothesis is outside the range of *schl, agep, eng,* and *gender,* we can confidently conclude that there is only a 5% chance the variable lies outside those boundaries. The null value lies inside the 95% interval for *racwht*, confirming its insignificance. These statistical conclusions are represented in Figure 1.12.

| **Independent Variable** | **Model 3** |
|---|---|
| *schl* | 0.062***<br>(0.003) |
| *agep* | 0.028***<br>(0.001) |
| *eng* | -0.064***<br>(0.018) |
| *racwht* | 0.026<br>(0.028) |
| *gender* | -0.487***<br>(0.028) |
| Intercept | 8.232<br>(0.084) |
| No. of obs. | 6020 |

| | |
|---|---|
| R-square | 0.213 |

*Figure 14; Source is Figure 13 in Appendix A; Statistical Significance: 1%***, 5%**, 10%**

**VI. Conclusions**

Our goal in conducting this research was to test the effect of educational attainment on individual wages using data from the state of Georgia. In our research, we discovered that the effect of educational attainment was economically and statistically significant at all levels. However, when measuring against other variables, educational attainment was not the greatest determinant of wages. Also, the addition of the dummy variable for gender had a greater effect on wages than any other variable in Model 3. In conclusion, within the scope of our project, gender had the greatest effect on wages and allowed for a greater explanation of the model. However, it should be noted that the increase in the R-squared value from Model 2 to Model 3 was not as high as expected, at 4%. In all, despite less significant economic interpretations, each variable in question, except for race, proved to be statistically significant at the 1% level. Going forward, we would recommend testing race for joint significance with other variables and if it was still found insignificant, ultimately dropping the race variable.

The societal implications of our results are not as positive as hoped, given gender had a greater effect on wage than schooling. Whereas we originally hypothesized that educational attainment, an aspect of choice, would be the greatest determiner of wages, our discovery leads us to believe that perhaps individuals have less control over their wages than they may want. However, this qualitative conclusion should be held loosely until the explanative scope of the model increases. To increase the explanative scope and decrease the omitted variable bias in future models, researchers should account for other variables currently accounted for in the error term. We advocate for variables accounting for experience or upbringing. Also, researchers should account for the educational shift currently taking place thanks to the technological revolution, as many educational programs for high paid positions, like software engineering, are conducted in informal, online, and unreported settings.

**Reference List**

"2017 ACS PUMS DATA DICTIONARY." ACS PUMS, October 18, 2018. https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2017.pdf?

Card, David. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica* 69, no. 5 (2001): 1127–60. https://doi.org/10.1111/1468-0262.00237.

Patrinos, Harry. "Estimating the Return to Schooling Using the Mincer Equation." *IZA World of Labor*, 2016. https://doi.org/10.15185/izawol.278.

Psacharopoulos, George, and Harry Anthony Patrinos. "Returns to Investment in Education: a Decennial Review of the Global Literature." *Education Economics* 26, no. 5 (July 2018): 445–58. https://doi.org/10.1080/09645292.2018.1484426.

Psacharopoulos, George, and Harry Anthony Patrinos *. "Returns to Investment in Education: a Further Update." *Education Economics* 12, no. 2 (2004): 111–34. https://doi.org/10.1080/0964529042000239140.

## Appendix A Stata Output

Figure 2 - a summary of each individual variable

```
. sum logwagp
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| logwagp | 47,795 | 10.18699 | 1.288478 | 1.386294 | 13.07946 |

```
. sum schl
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| schl | 96,647 | 15.94934 | 5.627567 | 1 | 24 |

```
. sum agep
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| agep | 99,799 | 40.32088 | 23.20507 | 0 | 92 |

```
. sum eng
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| eng | 11,082 | 1.601606 | .8812792 | 1 | 4 |

```
.
```

Figure 3 - correlation between *agep* and *logwagp*

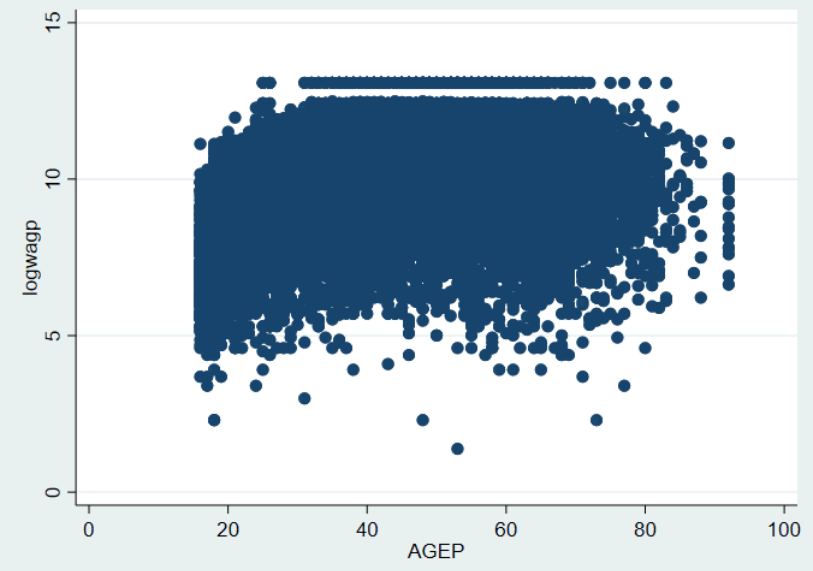

Figure 4 - correlation between *schl* and *logwagp*

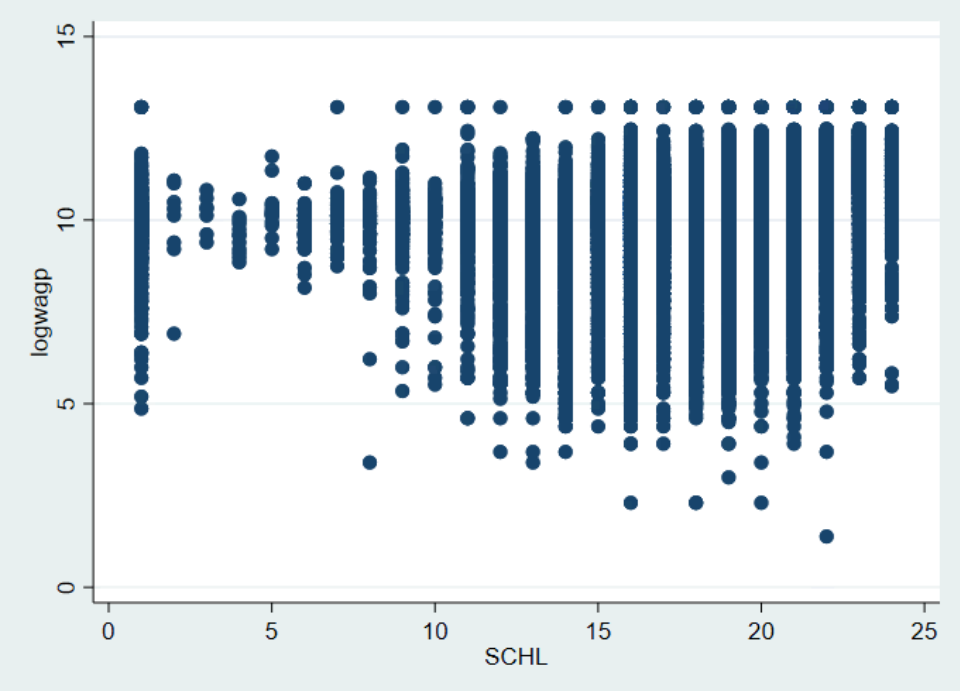


Figure 5 - correlation between *eng* and *logwagp*

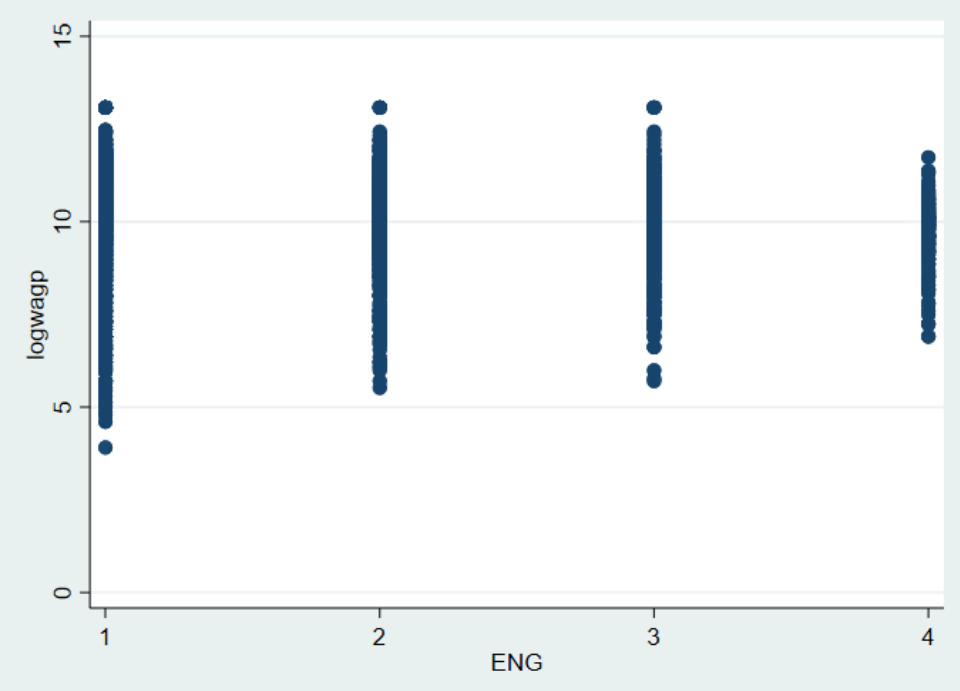

Figure 6 - correlation matrix output

```
. correlate logwagp schl agep eng
(obs=6,020)
```

| | logwagp | schl | agep | eng |
|---|---|---|---|---|
| logwagp | 1.0000 | | | |
| schl | 0.2702 | 1.0000 | | |
| agep | 0.3291 | 0.0570 | 1.0000 | |
| eng | -0.1200 | -0.4826 | 0.1083 | 1.0000 |

Figure 7 - simple linear regression output (Model 1)

```
. regress logwagp schl
```

| Source | SS | df | MS | Number of obs | = | 47,795 |
|---|---|---|---|---|---|---|
| | | | | F(1, 47793) | = | 5592.06 |
| Model | 8311.4922 | 1 | 8311.4922 | Prob > F | = | 0.0000 |
| Residual | 71034.9021 | 47,793 | 1.48630348 | R-squared | = | 0.1047 |
| | | | | Adj R-squared | = | 0.1047 |
| Total | 79346.3943 | 47,794 | 1.6601748 | Root MSE | = | 1.2191 |

| logwagp | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| schl | .1263739 | .0016899 | 74.78 | 0.000 | .1230616 | .1296863 |
| _cons | 7.845374 | .031806 | 246.66 | 0.000 | 7.783034 | 7.907715 |

Figure 8 - multiple linear regression output (Model 2)

```
. regress logwagp schl agep eng
```

| Source | SS | df | MS | Number of obs | = | 6,020 |
|---|---|---|---|---|---|---|
| | | | | F(3, 6016) | = | 419.77 |
| Model | 1510.23066 | 3 | 503.410219 | Prob > F | = | 0.0000 |
| Residual | 7214.72672 | 6,016 | 1.19925644 | R-squared | = | 0.1731 |
| | | | | Adj R-squared | = | 0.1727 |
| Total | 8724.95737 | 6,019 | 1.44956926 | Root MSE | = | 1.0951 |

| logwagp | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| schl | .0574082 | .0033527 | 17.12 | 0.000 | .0508358 | .0639806 |
| agep | .0285853 | .00106 | 26.97 | 0.000 | .0265074 | .0306631 |
| eng | -.0602084 | .0188877 | -3.19 | 0.001 | -.097235 | -.0231818 |
| _cons | 8.096401 | .0836532 | 96.79 | 0.000 | 7.932411 | 8.260392 |

Figure 13 - multiple linear regression output (Model 3)

```
. regress logwagp schl agep eng racwht gender
```

| Source | SS | df | MS | Number of obs | = | 6,020 |
|---|---|---|---|---|---|---|
| | | | | F(5, 6014) | = | 325.46 |
| Model | 1858.08168 | 5 | 371.616335 | Prob > F | = | 0.0000 |
| Residual | 6866.8757 | 6,014 | 1.14181505 | R-squared | = | 0.2130 |
| | | | | Adj R-squared | = | 0.2123 |
| Total | 8724.95737 | 6,019 | 1.44956926 | Root MSE | = | 1.0686 |

| logwagp | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| schl | .0619863 | .0032887 | 18.85 | 0.000 | .0555392 | .0684334 |
| agep | .028368 | .0010344 | 27.42 | 0.000 | .0263403 | .0303958 |
| eng | -.0639604 | .0184496 | -3.47 | 0.001 | -.1001283 | -.0277926 |
| racwht | .0258279 | .0276225 | 0.94 | 0.350 | -.0283221 | .0799779 |
| gender | -.4856979 | .0278801 | -17.42 | 0.000 | -.5403529 | -.4310429 |
| _cons | 8.231926 | .0843546 | 97.59 | 0.000 | 8.066561 | 8.397291 |